Proceedings of Statistics Canada Symposium 2001
Achieving Data Quality in a Statistical Agency: A Methodological Perspective

EDITING AND IMPUTATION IN A STANDARD ECONOMIC PROCESSING SYSTEM

Richard Sigman[1]

ABSTRACT

The U.S. Census Bureau developed software called the Standardized Economic Processing System (StEPS) to replace 16 separate systems used to process over 100 current economic surveys. This paper describes the methodology and design of the StEPS modules for editing and imputation and summarizes the reactions of users to using these modules to process their surveys.

Key Words: Survey Processing, Economic Surveys, StEPS

1. INTRODUCTION

The U.S. Census Bureau conducts surveys of households, institutions, and businesses. Surveys of businesses are the focus of this paper. The Census Bureau refers to these surveys as *economic surveys* because they provide economists and other analysts with estimates and data sets needed for macro- and micro-economic analyses. For example, the Bureau of Economic Analysis uses estimates from economic surveys to determine the national income and expense accounts. Economic surveys can differ widely with respect to characteristics of reporting units and content of survey questions. They are often similar, however, with respect to data-processing requirements, which prompted the Census Bureau to consolidate the survey-processing systems for many of its economic surveys. The development and use of generalized software, called the Standard Economic Processing System (StEPS), has made this possible.

This paper describes the editing and imputation capabilities of StEPS. Section 2 chronicles the development of the editing and imputation modules, and Section 3 briefly describes the entire StEPS system. Section 4 describes in detail the editing and imputation modules. Section 5 present two examples of data stored in the StEPS system that have been used to produce quantitative survey management information. Section 6 summarizes user feedback about using StEPS to perform editing and imputation, and Section 7 presents conclusions and describes future activities.

2. DEVELOPMENT OF THE StEPS EDITING AND IMPUTATION MODULES

The Census Bureau consists of several directorates that conduct censuses and surveys. The Economic Programs Directorate, conducts economic censuses every five years and conducts current economic surveys monthly, quarterly, and annually in areas of manufacturing, construction, commercial services, government services, and foreign trade. The Census Bureau directorates are responsible for developing survey methods and associated processing systems for the censuses and surveys they conduct.

The Economic Programs Directorate developed StEPS to process its current surveys. Though a few of these surveys allow respondents to provide data over the internet, these surveys are primarily mail surveys with telephone follow-up. Many of these surveys have subject-matter analysts perform telephone follow-up, but a few have clerks located in a centralized calling facility that contact mail nonrespondents. Current economic surveys are diverse in terms of size, frequency, and type of reporting units. The frequencies of current

[1]Richard Sigman, U.S. Census Bureau. ESMPD, Room 3108-4, U.S. Census Bureau, Washington DC 20233, USA, richard.s.sigman@census.gov . This paper reports the results of research and analysis undertaken by Census Bureau staff. It has undergone a more limited review by the Census Bureau than its official publications. This report is released to inform parties and to encourage discussion.

economic surveys include monthly, quarterly, annually, and quadrenially. The sample sizes of the surveys currently being processed by StEPS range in size from 12 for the Glass Containers Survey to 60,000 for the Annual Capital Expenditures Survey. Reporting units for current economic surveys include business establishments, divisions of companies, entire companies, and construction projects.

The development of StEPS started in 1995, with the gathering of user requirements and considering different design approaches. Interviews with survey managers, survey-processing specialists, and system designers were conducted and planning documents written by StEPS developers were reviewed by "advisory consultants", who were knowledgeable of the processing needs of the initial surveys that would be using StEPS (Ahmend and Tasky, 2001).

The determination of user requirements and possible design approaches was greatly assisted by three earlier activities. One of these was that prior to developing STEPS, the Economic Directorate had developed two other processing systems capable of processing more than one survey. The Current Industrial Reports (CIR) system was developed in the late 1980's for processing 75 surveys. These surveys collect data from manufacturers of industrial products; such as paint, light-bulbs, and iron and steel castings. The CIR system was written in FORTRAN and its edit and imputation operations were controlled by survey-specific parameters, which experienced users were able to specify by crating ASCII files containing fixed-field parameter "cards". The Generalized Annual Survey Processing (GASP) system was developed in the early 1990's for processing eight annual surveys that collected data from business firms involved in trade and services. The GASP system was written in COBOL, and its edit and imputation operations were controlled by survey-specific parameter files prepared by programmers. The edit parameter files contained COBOL-like text, which was converted to a COBOL "include" file. The imputation parameter files contained statements written in a custom-developed post-fix language, which was interpreted by the imputation software. The strength of the GASP system was its capabilities to estimate model parameters needed for model-based imputation functions.

Another earlier activity, which provided useful information about editing and imputation requirements, was the compilation by King and Kornbau (1994) of an inventory of the Economic Programs Directorate's statistical practices. King and Kornbau found that the editing and imputation practices of current economic surveys vary considerably in terms of the involvement of clerks, statistical assistants, and subject-matter analysts. For some of these surveys, edit checking and error correction are completely automated with, in some cases, referral of unusual units to subject-matter analysts for manual review. For other current surveys only edit checking is automated, and error correction is done by clerks or subject-matter analysts. In even other surveys, edit checking is only partially automated. King and Kornbau reported that many of the survey analysts they interviewed wanted to increase the amount of automation present in their edit and imputation procedures and some survey managers were interested in incorporating the use of graphics into their editing systems, such as using scatter plots to identify outliers.

The third earlier activity, which provided useful information about designing an editing and imputation system, was the development in 1994 of a processing system for the Farm and Ranch Irrigation Survey (FRIS). This system was written SAS® and many of its features for interactively specifying edit tests and reviewing edit results were incorporated into StEP. These features included the following (Monahan, 1994a, 1994b):

- The user interactively defined edit tests by using menus of common tests (presence tests, range tests, balance tests, verification tests, etc) and using SAS expressions to define more involved tests .
- The system applied the defined edit tests to the survey data and generated lists of edit-failing records, along with information about the items associated with each failed edit.
- An interactive review-and-correction module allowed the user to examine data in edit-failing records. The user could change data and mark data fields for imputation. The review-and-correction module was also used to input data values for variables that were previously unreported. An audit file recorded all changes to the data and permitted reversing of changes, if necessary.

At the same time that the Economic Programs Directorate was developing StEPS, it was also developing an edit and imputation subsystem, called Plain Vanilla, for processing its 1997 economic census. Plain Vanilla (PV) is so named because it provides general-purpose editing and imputation capabilities that one can augment with survey-specific computer code (i.e. "toppings"). Because of its high data volumes, the economic census

automates more of its editing and imputation process than the current economic surveys do. The economic census limits its types of edits to ratio edits, balance edits, and survey-code-verification edits; performs error localization of ratio-edit failures; and has subject-matter analysts examine only referred cases. The development of StEPS edit and imputation modules and the development of PV were similar in that user requirements arose from users' experiences with earlier systems (the CIR, GASP, and FRIS systems for StEPS; and the SPEER system for PV) and the primary development objective for both was replacing multiple editing and imputation systems with a single system. These development activities differed, however, in that they developed different types of systems–a highly automated system for the economic census and a very flexible and easily configured system for the current economic surveys. Nevertheless, there was a beneficial knowledge transfer from PV development to StEPS development in the area of imputation methods for data failing balance edits (Sigman and Wagner, 1997). Much of the PV development work in this area was implemented in StEPS.

3. OVERVIEW OF StEPS

StEPS is a generalized survey processing system that the Economic Directorate developed to replace 16 legacy systems. In addition to reducing resources needed for system maintenance, one of the StEPS objectives is to shift more processing control to survey analysts and methologists. StEPS contains integrated modules for data-collection support (e.g., mail-label printing and questionnaire check-in); editing; data review and correction; imputation; calculation of estimates and variances; and system administration (e.g., parameter specification and the submission and monitoring of batch jobs). Functions not in StEPS include frame development, sample selection, actual data collection, and data dissemination. StEPS is programmed in SAS, and it stores data and parameters in SAS data sets. The Economic Directorate executes StEPS mainly on Compaq® Alpha® machines using UNIX as the operating system. Most users access StEPS via a graphical (X-Windows) communication package loaded on their desktop microcomputer.

Ahmed and Tasky (1999, 2000, 2001) provide additional information about StEPS. Tasky et al. (1999) describe the StEPS system design and associated programming strategies. In particular, they state that the developers of StEPS "decided on four major design concepts:

1. "Design a set of standard data structures that remain the same, regardless of the survey and the data.
2. "Use parameters (stored in general data structures) to drive the survey-specific processing requirements.
3. "Generate a 'fat' record data set [containing all data for a reporting unit in one record] on the fly for certain modules ... .
4. "Standardize field names and possible values for similar concepts."

StEPS uses SAS variable names to refer to individual items of survey data. On the fat-record data sets, the naming convention for variable names is *trrrrrxx*, where *t* is the type of data (*t*='R' for reported data, *t*='E' for edited data, *t*='A' for adjusted data, and *t*='W' for weighted adjusted data), *rrrrr* is a root name or key code for the item (up to five characters in length), and *xx* is the relative statistical-period indicator (*xx*="00" for current statistical period, *xx*="01" for prior statistical period, etc). Since StEPS initially loads respondent data into both *Rrrrrrxx* (i.e. reported data) and into *Errrrrxx* (i.e. edited data), we will in the examples that follow use the edited-data item names when referring to survey data stored in StEPS.

4. StEPS EDITING AND IMPUTATION MODULES

StEPS has a module for data editing and two modules for imputation. The two imputation modules perform what StEPS labels simple imputation and general imputation. The usual order in which the modules are executed is first simple imputation, then editing, and finally general imputation.

4.1. Simple Imputation Module

The StEPS simple-imputation module imputes data values considered to be equivalent to reported data. The resulting data are flagged as being reported. A frequently performed type of simple imputation is "data filling";

i.e., StEPS fills in data that the respondent failed to provide when the value the respondent should have provided can be easily inferred from other data. For example, the Annual Retail Survey (ARTS) collects data for

etaxyn00 = indicator for collection of sales tax by retailer (1 for "yes", 2 for "no"),
ectax00 = annual amount of sales tax collected,
ecsal00 = annual sales, excluding sales tax, and
ectsal00 = annual total sales, including sales tax.

If *etaxyn00* is missing (or equal to 1 for "yes") but *ectax00*=0 and *ecsal00*=*ectsal00*>0, then one of the StEPS simple-imputation rules for ARTS sets *etaxyn00* to 2 (for "no").

StEPS users can interactively specify two types of simple-imputation rules: balance complex rules and free-form rules. For the latter, the user specifies SAS expressions that describe "error" conditions and corresponding actions to be taken when the conditions are satisfied. The specified actions can be any group of SAS statements, regardless of their complexity. Balance-complex rules are associated with additive relationships between a total, denoted y, and details denoted x_i, i=1,2,...n. The user specifies one more adjustments to y and/or to the x_i to be performed when $y \neq \Sigma x_i$, y is missing, and/or one or more x_i is missing, and the data that are available for the balance complex are complete enough for imputed values to be considered equivalent to reported data. The completeness of the available data is determined by testing if the absolute residual $|R|=|y-\Sigma x_i|$ and/or the relative absolute residual $|R|/y$ are less than user-specified tolerance values. The user can specify one or more of the following adjustment actions to be performed (Luery, 1999):

ZERO-SET. Set missing x_i to zero.
YSUMX. Set y to $\Sigma_{nm} x_i$, where $\Sigma_{nm} x_i$ is the sum of the non-missing x_i.
RESIDUAL. If only one x_i is missing, set it to $R=y-\Sigma_{nm} x_i$.
RAKE. Calculate adjusted x_i, denoted x_i', such that $y=\Sigma x_i'$.

When y and all the x_i are non-negative, the RAKE option uses the formula $x_i' = x_i(y/\Sigma x_i) = x_i(1+R/\Sigma x_i)$, which is well known to subject-matter analysts as an acceptable practice when $|R|/y$ is small, say less than 0.05. This practice has a sound statistical basis in situations in which the error in reporting a detail, x_i, occurs at random and the variance of the random error in reporting x_i, denoted $var(x_i)$, is proportional to x_i. Then using the method of Lagrange multipliers, it can be shown that the raked details, x_i', minimize the chi-square statistic

$$\chi^2 = \sum_i \frac{(x_i' - x_i)^2}{var(x_i)}$$

subject to the constraint $y=\sum x_i$ (See Deming, 1943, Chapter 5.). When the x_i are not restricted to being non-negative, Luery and Sigman (2000) show that if $var(x_i)$ is proportional to $|x_i|$ then minimizing the above chi-square statistic subject to $y=\Sigma x_i'$ yields the adjustment formula $x_i' = x_i\,[1+\text{sign}(x_i)\,R/\Sigma|x_i|]$.

Interactive screens permit users to specify simple-imputation rules. These rules can then be executed in batch for all cases or can be executed interactively in the StEPS review-and-correction module for a single case.

4.2. Editing Module

The StEPS editing module performs automated detection of possible data errors, which are data values that individually or in relationship to other data fail to conform to expected reporting behavior. The StEPS editing module only identifies the failures; it does not change data. The editing module also allows users to interactively define edits and to examine edit results in a variety of ways. Though we do not discuss it further in this paper, the editing module is currently being enhanced to include the statistical editing approach described by Hidiroglou and Berthelot (1986) and evaluated by Hunt, et al. (1999) on data from the Monthly Retail Trade Survey.

Users can define the following type of StEPS edits (Tasky, 2000a):

Required-data-item test. Verifies that the value of a specified item is not equal to missing.
Range test. Verifies item value lies in the range defined by specified minimum and maximum values.
List-directed test. Verifies the value of the specified item is contained in a pre-defined list of values.
Balance test. Verifies that a sum of specified detail items is equal to a specified total.

Survey-rule test. Free-form test that validates complex inter-item relationships.
Negative test. Verifies that the value of the specified item is not negative.

The list-directed test and the survey-rule test are conditional tests–they have a pre-condition that must be satisfied before the edit condition is tested. The other StEPS edit tests do not have pre-conditions. Conditional edits are commonly used when sample units are selected from different economic sectors and data from different sectors are to be edited differently. Because of its pre-condition and its flexibility, the survey-rule test is the most frequently used edit test. (Section 5, below, provides additional information about the frequency of use of different StEPS editing and imputation options.) The following are some examples of edit-rule tests for ARTS, involving items *ectax00* (annual sales tax collections), *ecsal00* (annual sales, excluding sales tax), *ectsal00* (annual total sales, including taxes), and the symbol "**.**" indicating missing data:

Only taxes reported. Pre-condition: none. Error condition: *ectax00*>0 and *ecsal00*=. and *ectsal00*=.
Taxes too large. Pre-condition: *ecsal00*≠**.** and ectsal00>0.
Error condition: *ectax00*>*ecsal00**0.15 or *ectax00*>*ectsal00**0.15

The definition of each edit includes a designation of how the edit will be used. The possible choices are one or more of the following usages (called "events" in StEPS):

Pre-edit. Designates a subset of the edits to be executed for all cases and the failures to be reviewed.
Full-edit. Designates the complete set of edits to be executed for all cases and the failures to be reviewed. (Negative tests are not included in the full edit.)
General-imputation edit. Designates edits to be executed for all cases to mark cases for imputation.
Single-ID edit. Designates edits to be executed in review-and correction module for an individual case.

When StEPS executes edits in batch that have pre-edit or full-edit event flags, information about the cases and items that failed the edits is put in a survey-level reject file. When these same edits are executed interactively, the edit-failure information is put in a user-defined reject file. These reject files may be examined in a variety of ways. They can be printed out or viewed online in a data-listing format. The reject files also support interactive editing in the StEPS review-and-correction modules, where one can view all the data associated with a case, view the corresponding edit failures in the reject files, change data in the case, and execute for the particular case the edits that have single-ID event flags. Also, in the review and correction module one can designate that a case be bypassed in subsequent editing runs. The use of the StEPS review-and-correction module to perform interactive editing allows subject-matter analysts to quickly correct detected data errors. Willimack et al. (2000) used analyst focus groups to determine that experienced analysts perform interactive editing in the following way:

"1. Review all edit failure messages.
"2. Identify 'easy' edit failures and resolve them.
"3. Resubmit the case to the ... edit ....to identify remaining errors.
"4. Characterize the case and, if necessary, do further research.
"5. Resolve any remaining edit failures.
"6. Iterate steps 3-5, as necessary.
"7. Call the company about unresolved edit failures."

4.3. General Imputation Module

Unlike simple imputation, the values changed by the general-imputation module are flagged as imputed data. The general-imputation module imputes data using estimator type techniques (Giles and Patrick, 1986) and adjusts balance complexes so that detail items sum to total items (discussed in Sigman and Wagner, 1997). Interactive screens allow users to select from menus of methods for imputing individual items and from menus of actions for adjusting balance complexes. Table 1 summarizes the methods for imputing individual items, using the following notation:

v = the item-name for the value being imputed
v' = the imputed value of v
z_j = the imputed value of the j^{th} auxiliary variable. (An auxiliary variable is a constant, an item name other than v, or an item-name/constant expression associated with the case for v is being imputed.)
$S(f)$ = the sum of item name f over a defined set of records

$(S(f_1)/S(f_2))_I$ = the ratio of identicals of item name f_1 to item name f_2. This is the ratio of two sums, both of which use all the records in an associated imputation cell in which both f_1 and f_2 are nonmissing and satisfy certain acceptance criteria. An example of the latter is that $L \le f_1/f_2 \le U$, where L and U are specified by the user.

Table 1. Methods for imputing individual items (Luery, 2001)

Name	Description	Formula
VALUE	Value of an auxiliary variable.	$v' = z_1$
SUM	Sum of auxiliary variables.	$v' = z_1 + z_2 + \ldots + z_n$
PRODUCT	Product of two auxiliary variables.	$v' = z_1 z_2$
RESIDUA	Auxiliary variable minus the sum of other auxiliary variables.	$v' = z_1 - (z_2 + \ldots + z_n)$
ATREND	Auxiliary variable multiplied by a trend.	$v' = z_1(z_2 / z_3)$
MEAN	Mean of an auxiliary variable over all records in an imputation cell satisfying certain acceptance criteria.	$v' = \bar{z}$
RATIO	Ratio prediction for imputed item .	$v' = (s(v) / S(z_1))_I z_1$
AUXRAT	Auxiliary variable times a ratio of identicals.	$v' = z_1(S(z_2) / S(z_3))_I$
SIMPREG	Auxilary variable times a regression coefficient.	$v' = b_1 z_1$
MULTREG	Multiple-regression prediction for imputed item..	$v' = b_1 z_1 + \ldots + b_n z_n$

When more than one method is selected to impute an item, the user specifies an order for StEPS to use to process the selected methods. Also, the user can assign to each selected method an imputation condition, which must be satisfied in order for StEPS to use the method. Table 2 contains the imputation specifications for the ARTS item *ectax00* (annual collected sales taxes), involving the following items:

ecsal00 = unweighted annual sales, excluding sales tax

etaxyn00 = indicator for sales-tax collection: 1 for "yes", 2 for "no"

wctaxy00 = recoded item that is equal to *wctax00* (weighted annual sales tax) when *etaxyn00*=1 and is missing otherwise

wctaxb00 = recoded item that is equal to *wctax00* when *etaxn00* is in {1,2} and is missing otherwise

Table 2. General-Imputation specifications for ARTS item *ectax00* (Burton, 2000)

Condition	Method	Formula	Auxiliary variables
etaxyn00=1	AUXRAT	$ecsal00*(S(wctaxy00)/S(wcsal00))_I$	z_1=*ecsal00*, z_2=*wctaxy00*, z_3=*wcsal00*
etaxyn00=.	AUXRAT	$ecsal00*(S(wctaxb00)/S(wcsal00))_I$	z_1=*ecsal00*, z_2=*wctaxb00*, z_3=*wcsal00*

For records in which *ectax00* is marked for imputation and *ectaxyn00*=1 (indicating collection of sales tax) the imputation of *ectax00* is based on a weighted ratio of identicals calculated from other records in the imputation cell that have *ectaxyn00*=1. For records in which *ectax00* is marked for imputation and *ectaxyn00* is missing, however, the imputation of *ectax00* is based on a weighted ratio of identicals calculated from records with either *ectaxyn00*=1 or *ectaxyn00*=2.

The available actions for adjusting balance complexes (involving details, denoted x_i , that sum to a total, denoted y) include the simple-imputation actions (ZERO_SET, YSUMX, RESIDUAL, and RAKE, defined in 4.1), plus the following ones (Luery, 2001, Section II):

RAKEIMP. Rake all previously imputed details. ($\sum_{imp} x_i$ and y-$\sum_{imp} x_i$ replace $\sum x_i$ and y-$\sum x_i$, respectively, in the RAKE formula, where $\sum_{imp} x_i$ is the sum of previously imputed details.)
ROUND. Divide details by 1000 and then rake. (x_i / 1000 replaces x_i in RAKE formula.)
NSK. Set a Not-Specified-by-Kind (NSK) variable to equal the residual $R= y$-$\sum x_i$ or add R to a specified detail or to the largest detail.

The calculation of imputed data is a batch process. The first step is the creation of a "fat" file containing all the survey variables. This file is passed through three times. The first pass marks data for imputation by executing general-imputation edits tests and by testing defined balance complexes. Additional control over the first-pass processing is provided by record-level bypass flags and item-level marked-for-imputed flags, both of which can be set in the review-and correction module. The second pass calculates needed means and ratios of identicals. Similar to the first pass, user-set flags at the record and item levels provide additional control over the exclusion of extreme or suspicious data from these calculations. The third pass imputes the data marked for imputation, retests balance complexes, and performs balance-complex adjustment actions. Like the first and second passes, additional control over this operation is provide by user-set record-level bypass flags and item-level marked-for-imputation flags (Tasky, 2000b).

5. SURVEY MANAGEMENT INFORMATION FROM StEPS FILES

Currently, 94 of the surveys conducted by the Economic Directorate use the StEPS editing and imputation modules. Each survey uses interactive screens in StEPS to create survey-specific editing and imputation rules, which include specifications for simple imputation, edit tests for identifying cases to be reviewed, edit tests for marking cases to be imputed, definitions of balance-complexes, and selections of methods for general-imputation of individual items. These survey-specific editing and imputation rules are stored in SAS data sets, and they customize StEPS to the particular data items and data relationships associated with each survey. These SAS data sets, containing survey-specific editing and imputation rules, can be analyzed (outside of StEPS) to yield information on how surveys are using the editing and imputation modules.

Table 3 contains counts of the editing and imputation rules for 75 surveys currently using StEPS, disaggregated by survey frequency and rule type. A complete discussion of Table 3 is beyond the scope of this paper, but one item of note is that the current use of general imputation is primarily for imputing missing data--i.e. imputing data marked for imputation by required-item tests–as opposed to correcting data rejected by edit tests other than required item tests. In fact, additional analysis of the 172 survey-rule tests that mark data for general imputation indicates that these 172 rules are associated with only four surveys. This suggests that for many of the surveys currently using StEPS there is no need for error localization since error localization requires that at least one of the failed edits involves more than one item.

StEPS edit tests are also used to identify data to be reviewed in the StEPS review-and-correction module. Here, data can be changed interactively by the user, and these changes are recorded in an audit trail. The audit trail is a SAS data set, and it can be analyzed (outside of StEPS) to provide quantitative information about intaractive editing. Farrar (2000) analyzed the StEPS audit trail for the 1998 Annual Trade Survey to study analyst changes to annual wholesale sales data. Some of his findings were the following:

- In 8.5 percent of the cases, annual sales data were interactively edited.
- In only eight percent of the analyst changes to annual sales data was a cased edited more than once, and only one case was corrected more than four times.
- "Sixty-five percent of the cases changed [by analysts changing annual sales data] involved cumulative chages greater than 100 million dollars (either positive or negative). Additionally, in some SICs the effect of [analysts'] edits on the final, published figures was dramatic. ... This indicates that manual analyst edits are focusing on the largest and most significant errors."

Table 3. Counts of editing and imputatation rules

	All surveys	Monthly surveys	Quarterly surveys	Annual Surveys	Quadrennial Survey

Number of surveys	75	9	12	53	1
Total number of questionnaire items	15,204	751	2900	10,962	591
Total number of edit and imputation rules:					
All types	54,782	3,108	10,649	40,649	249
Simple imputation:					
Free form	138	0	0	138	0
Balancing	101	0	0	101	0
Edit tests:					
For review	18,635	990	4,502	12,894	249
For general imputation	14,127	739	2,900	10,448	0
General imputation rules:					
Balancing complexes	846	21	95	730	0
Methods for item imputes	21,005	1,430	3,152	16,423	0
Breakdown of number of edit-tests:					
For review:					
All types	18,635	990	4,502	12,894	249
Required-item tests	2	0	0	1	1
Balance tests	2,246	44	675	1505	22
Survey-rule tests	16,381	946	3,827	11,382	226
Other types	6	0	0	6	0
For general imputation					
All types	14,127	739	2,900	10,488	0
Required-item tests	13,955	739	2,900	10,316	0
Survey-rule tests	172	0	0	172	0
Breakdown of number of general-imputation methods for item imputes.					
All methods	21,005	1,430	3,152	16,423	0
RATIO	7,856	671	959	6,226	0
AUXRAT	2,468	36	432	2,000	0
VALUE	10,475	723	1761	7,991	0
ATREND	194	0	0	194	0
SIMPREG, SUM, PRODUCT	12	0	0	12	0

6. USER FEEDBACK ABOUT STEPS EDITING AND IMPUTATION

Surveys in the Economic Directorate currently using StEPS used other systems prior to StEPS. These earlier systems were very different from StEPS--they were survey specific, to add or remove survey items they often required changes in computer code, and often only programmers were able to change parameters and submit production jobs. The operational differences between these earlier systems and StEPS has required paradigm shifts by managers, methodologists, and analysts. At an October 2000 methodological interchange between Statistics Canada and the U.S. Census Bureau, a panel of StEPS users discussed what they liked and disliked about the the StEPS editing and imputation modules (Burton and Hanks, 2000). Some of the things the panel members liked about the editing and imputation modules were the following:

- StEPS is a repository for accepted statistical methods.
- It is easy to select methods in general imputation.
- The imputation and balancing methods are well documented.
- Implementation problems are easy to solve by those who know SAS.
- Batch job streams are easy to understand by those who know SAS.
- The survey-rule edit test is very flexible.
- Different types of edits can be executed separately.
- Edits can be run for a single observation or a subset of observations.
- Edit-failing cases can be grouped together in a file and then reviewed.
- Specification screens allow analysts to control the survey processing.
- Analysts can submit their own jobs.

- Separate components of the imputation process can be individually tested.
- Results form StEPS are similar to those from earlier systems.

Some of the things the panel members did not like about the StEPS editing and imputation modules were the following:

- The run times for some batch jobs are quite long, though progress has been made in reducing run times.
- There is a steep learning curve, especially for users that don't know SAS.
- The interaction between different parts of StEPS and with survey requirements can be complex. Sometimes SAS programs must be written to find subtle implementation problems.
- Many parameters are required and creating and maintaining them can be difficult.
- Balance-complex definitions have to be specified in both the editing module and in the imputation module.
- Imputation methods that calculate ratios of identicals are not easy for analysts to implement.

One panel member mentioned that the following features of the StEPS editing an imputation modules were both positive and negative:

- The flexibility of StEPS empowers the user, but using StEPS can be user intensive.
- SAS files and SAS syntax makes StEPS a very flexible system for those who know SAS but can make it a more difficult system to use for those who do not know SAS.
- Detailed user-set flags control StEPS editing and imputation, and as a result these operations can be complex.

7. CONCLUSIONS AND FUTURE ACTIVITIES

StEPS offers a flexible approach to editing and imputation. It provides capabilities for data filling, interactive editing (i.e. edit tests, plus interactive online correction and retesting), statistical edits, and machine imputation. The StEPS edit and imputation modules are configured to different surveys through interactive screens that allow users to define edit and imputation rules and to submit their own jobs to evaluate rules they have defined. StEPS stores data, data changes, and processing parameters in SAS data sets, which permits SAS-knowledgeable survey practitioners to develop quantitative survey-management information. Users of the StEPS editing and imputation modules like the flexibility and empowerment StEPS provides, but they note that some of the disadvantages of generalized systems are longer completion times for batch processes, increased complexity in the relationships between different processing activities, and a steep learning curve in becoming proficient in configuring StEPS to different survey situations.

The Economic Directorate plans to increase the number of its surveys that use StEPS. Planned enhancements to the StEPS editing and imputation modules include the addition of hot-deck (or nearest neighbor) imputation, improvements to graphical editing, and the addition of capabilities for macro review of tabulated results.

REFERENCES

Ahmed, S. and Tasky, D. (1999), "The Standard Economic Processing System: A Generalized Integrated System for Survey Processing, " *Proceedings of the Section on Government Statistics and Section on Social Statistics*, American Statistical Association, pp. 205-210.

_____ (2000), "Standardized Economic Processing System," *Proceedings of the International Conference on Establishment Surveys*, Alexandria, VA: American Statistical Association, pp 633-642.

_____ (2001), "Are Generalized Systems the Way of the Furure: A Case Study on the Standard Economic Processing System (StEPS)," *Proceedings of the Survey Research Methods Section*, Alexandria VA: American Statistical Association, to appear.

Burton, J. (2001). "General Imputation Memo for StEPS: Supplement 8 (ARTS)," unpublished memorandum, Washington DC: U.S. Census Bureau, Services Sector Statistics Division.

Burton, J. and R. Hanks (2000). Panel discussion, Session 8: Applications of Generalized Processing Systems, 2000 Statistics Canada/Census Bureau Methodological Interchange, Washington DC: U.S. Census Bureau, Methodology and Standards Directorate.

Deming, W.E. (1943). Statistical Adjustment of Data, New York: Wiley.

Farrar, R. (2000). "The StEPS Audit Tail File as a Survey Management Tool," unpublished report, Washington DC: U.S. Census Bureau, Economic Planning and Coordination Division.

Giles, P. and C. Patrick (1986). "Imputation Options in a Generalized Edit and Imputation System," Survey Methodology, v. 12, pp. 49-60.

Hidiroglou, M. and J. Berthelot (1986). "Statistical Editing and Imputation for Periodic Business Surveys," Survey Methodology, v 12, pp. 73-83.

Hunt, J; J. Johnson, and C. King (1999). "Detecting Outliers in the Monthly Retail Trade Survey Using the Hidiroglou-Bethelot Method, *Proceedings of the Survey Research Methods Section*, Alexandria VA: American Statistical Association, pp. 539-543.

King, C. and Kornbau, M. (1994), *Inventory of Economic Area Statistical Practice, Phase 2: Editing, Imputation, Estimation, and Variance Estimation*, Technical Report #ESMD-9401, Washington DC: Bureau of the Census, March 1994.

Luery, D. (1999). "Simple Imputation for One Dimensional Balance Complexes," StEPS Decision Document #10, Washington D.C.: Bureau of the Census, Economic Statistical Methods and Programming Division.

_____ (2001). "General Imputation," unpublished documentation, Washington DC: U.S. Census Bureau, Economic Statistical Methods and Programming Division.

Luery, D. and R. Sigman (2000). "Raking When the Details are Positive and Negative," unpublished documentation, Washington DC: U.S. Census Bureau, Economic Statistical Methods and Programming Division.

Monahan, J. (1994a). "Farm and Ranch Survey Data Processing Concepts," internal documentation, Washington, D.C.: U.S. Bureau of the Census, October 24, 1994.

_____ (1994b). "The FRIS Processing System," internal documentation, Washington, D.C: U.S. Bureau of the Census, November 20, 1994.

Sigman, R. and D. Wagner (1997), "Algorithms for Adjusting Survey Data That Fail Balance Edits,"*Proceedings of the Survey Research Methods Section*, Alexandria VA: American Statistical Association.

Tasky, D. (2000a). "StEPS Edit / Imputation Seminar - 2," unpublished documentation, Washington DC: U.S. Census Bureau, Economic Statistical Methods and Programming Division.

_____ (2000b). "Flow of General Imputation," unpublished documentation, Washington DC: U.S. Census Bureau, Economic Statistical Methods and Programming Division.

Tasky, D.; Linonis, A.; Ankers, S; Hallam, D., Altmayer, L.; and Chew, D. (1999). "Get in Step with StEPS: Standard Economic Processing System," *Proceedings of the North East SAS Users Group*, pp. 167-178.

Willimack, D.; A.E. Anderson, and K.J. Thompson (2000). "Using Focus Groups to Identify Analysts' Editing Strategies in an Economic Survey," *Proceedings of the International Conference on Establishment Surveys*, Alexandria, VA: American Statistical Association.